PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Extraction and Use of Facial Features in Low Bit-Rate Visual Communication

Don Pearson

Phil. Trans. R. Soc. Lond. B 1992 **335**, 79-85
doi: 10.1098/rstb.1992.0010

The extraction and use of facial features in low bit-rate visual communication

DON PEARSON

Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, U.K.

SUMMARY

A review is given of experimental investigations by the author and his collaborators into methods of extracting binary features from images of the face and hands. The aim of the research has been to enable deaf people to communicate by sign language over the telephone network. Other applications include model-based image coding and facial-recognition systems. The paper deals with the theoretical postulates underlying the successful experimental extraction of facial features. The basic philosophy has been to treat the face as an illuminated three-dimensional object and to identify features from characteristics of their Gaussian maps. It can be shown that in general a composite image operator linked to a directional-illumination estimator is required to accomplish this, although the latter can often be omitted in practice.

1. INTRODUCTION

Over the past decade the author has been involved in experimental investigations designed to extract binary representations of the face and hands from camera images. The aim of this research has been to allow deaf people to communicate over telephone lines by using sign language. Although the telephone network was designed to transmit voice signals, it can also carry data at rates up to about 14.4 kbits s^{-1}; this is just sufficient to enable small moving images to be transmitted. In devising coding techniques to realize such very low bit rates, feature extraction in line-drawing form was used as a source coding operation to reduce the information in the image.

In the course of these endeavours a good deal has been learnt about the economical representation of the face and hands; indeed, theoretical formulations have been a key factor in the success of the project. In this paper the main findings that relate to facial processing will be reviewed, including recent generalizations of the theory which deal with the direction of illumination and the role played by parabolic facial curves.

2. CODING IMAGES OF THE FACE AND HANDS INTO VERY LOW BIT RATES

Following earlier work with the Royal National Institute for the Deaf using broadband coaxial cable (Pearson & Sumner 1976), a study was made of the feasibility of coding moving images into telephone data rates (Pearson 1981). The results were encouraging and gave rise to research designed to simplify the visual representation of the face and hands. Initial attempts to achieve this using edge detection resulted in some success but with a cluttered representation of internal facial features (Pearson & Six 1982; Pearson 1983). In 1983 a new theory was formulated in three-dimensional object space for identifying key features; when implemented it resulted in a startling improvement in the economy and verisimilitude of the representations (Robinson & Pearson 1984; Pearson & Robinson 1985; Robinson 1985, 1986; Pearson 1986). In 1984 two-way sign-language conversations were successfully conducted in the laboratory at a simulated rate of less than 10 kbits s^{-1}.

In 1986 collaboration with British Telecom Laboratories commenced in an attempt to realize a practical system. By this time techniques for transmitting data over telephone lines had improved and reliable full-duplex transmission at 14.4 kbits s^{-1} was possible. The prototype system that emerged (Whybray & Hanna 1989) used videophone hardware developed for another purpose; nevertheless, by the addition of interframe coding the necessary data rate was achieved. A successful two-way transmission over a public switched telephone connection was accomplished between the University of Essex at Colchester and British Telecom Laboratories at Martlesham on 6 February 1989, followed by the first-ever sign-language conversation on 15 February 1989 between the home of a deaf couple at Rushmere St. Andrew and the Suffolk Deaf Association in Ipswich. The prototype system has been in daily use ever since.

Important studies with a similar aim have been conducted elsewhere in Europe, in the U.S.A. and in Japan (Sperling *et al.* 1981, 1985; Letellier *et al.* 1985; Ono *et al.* 1985). It is believed, however, that the February 1989 transmissions were the first of their kind in the world. More recently a laboratory demonstration has been achieved at BTL of sign-language communication by using small moving grey-level images, coded by hybrid interframe techniques into

14.4 kbits s^{-1} (Whybray 1991). Such developments hold promise of further intelligibility improvements in the 1990s.

3. OPERATORS FOR FACIAL FEATURE EXTRACTION

The theory responsible for the successful extraction of facial features and of the outlines of the hands utilizes mappings of three-dimensional objects onto a Gaussian sphere. A comprehensive treatment of Gaussian maps has been given by Koenderink (1990); the principle is that any elemental surface area of a solid is mapped onto the Gaussian sphere according to the direction of its surface normal. Curved surfaces like the nose or fingers generate large areas on the Gaussian sphere, whereas flat regions such as the forehead generate small areas.

As initially formulated, the basic postulate of the theory was that the important surfaces to delineate in a line drawing or 'cartoon' are those which map to a small annular region on the Gaussian sphere, bounded on the outside by a great circle whose centre corresponds to the viewing direction (figure 1). We shall call this region the feature region or f-region. The postulate can be equivalently stated in other ways, for example, that such surfaces have a surface normal which is nearly perpendicular to a straight line drawn from the eye or camera, or that they are grazed by this straight line, or that they lie close to what Koenderink calls the rim (Pearson & Robinson 1985; Pearson *et al.* 1990). Note that the rim maps onto the great circle. The postulate on important facial surfaces was a purely intuitive one, based on inspection of the line drawings of artists as well as personal observation of faces and hands.

To use the postulate for feature extraction with digital images, it has to be translated into a series of operations in image space. Analysis showed that under general conditions of frontal illumination, matt objects of uniform reflectance seen against a similar background gave rise to luminance valleys at the surfaces specified by the postulate. Later it was realized that any valley detector with a finite region of support also has a response to large edges; in consequence the term 'valledge' detector was coined to describe the operator.

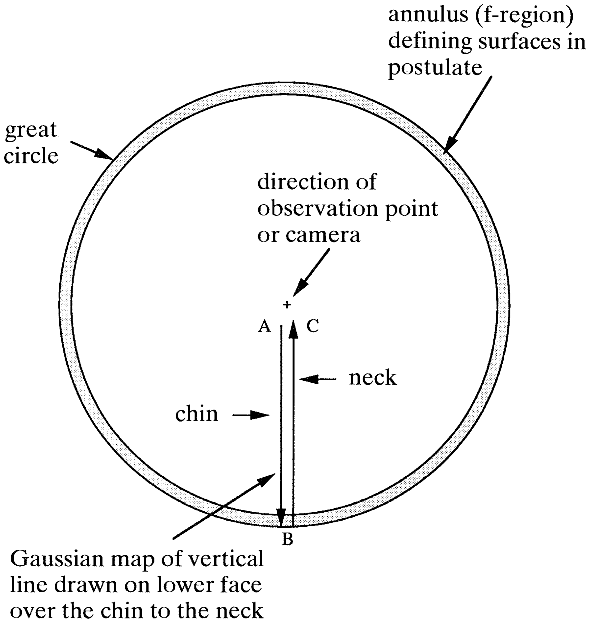

Figure 1. Gaussian sphere showing the f-region used for defining important facial surfaces, together with the map of line ABC in figure 2.

A further embellishment was added in the form of black shading. It is noticeable how effective this is in artist-drawn representations (figure 2*a*) and it is easily produced electronically by a thresholding operation. At the time we introduced it we had no scientific explanation for its effectiveness, although we did note that it eliminated low-luminance feature lines and therefore made the image easier to code. With the addition of thresholding the operator was three-composite, that is with a response to three features in the image: valleys, large edges and dark areas. The use of this three-composite operator on a grey-level image produces the representation shown in figure 2*b*. The same grey-level image was used by the artist as a basis for drawing figure 2*a*, so the two can be compared for feature lines.

We have reported comparisons between artist-drawn and machine-drawn cartoons in Pearson *et al.* (1990). Their remarkable similarity suggested a reason why a line drawing can convey so much information about a face, provided the lines are in the right place. This could be, we speculated, because early human vision might involve operations similar to those we had used in the generation of our machine-drawn cartoon. As we had established experimentally that a machine cartoon of a machine cartoon was itself, we thought that artists might unconsciously put lines in certain positions on a piece of paper because those lines produced a similar neural output in early human vision as did the original face.

In 1989 it was decided to disassemble the three-composite operator to find out what parts of the face were being detected by the valledge and thresholding components (Pearson & Hanna 1989). This study showed that the valledge component was good at delineating the nose, cheeks and chin, while the thresholding operation worked well in bringing out dark hair, eyebrows and any bounding surfaces seen against shadow, such as the nostrils, interior of the mouth, or chin. The valledge and thresholding operations tended to complement each other, this being particularly true in regions where the facial surface is complex, such as the eyes. More recently Bruce *et al.* (1991) have conducted an extensive study of the separate and combined effects of the two operators in facial-recognition experiments. They found that recognition performance is superior if both the valledge and thresholding components are used; with both components, recognition was almost as good as with grey-level pictures.

In addition to their possible application in visual communication and in facial-recognition systems, binary extracted facial features are employed in model-based image coding to locate and track the

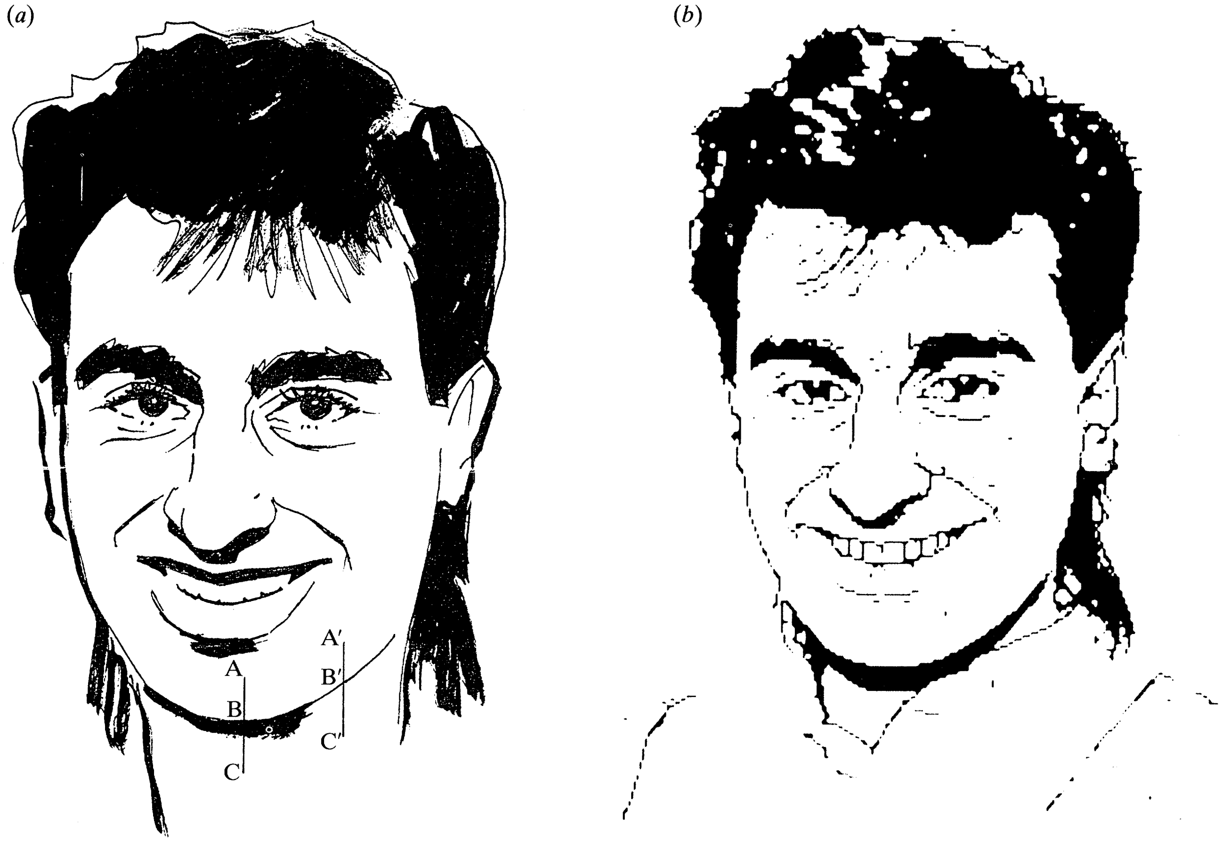

Figure 2. (*a*) Artist-drawn and (*b*) machine-drawn representations of the same face (from Pearson *et al.* 1990).

head (Welsh 1991). There is computational simplicity and consequently faster processing in using them instead of grey-level images. However, the results can depend critically on the operator used; too often an inappropriate choice is made by experimenters as a consequence of haste or a poor understanding of what the operator does. Bergen & Adelson (1988) have shown that getting the operator right in modelling early vision can make more complicated high-level explanations unnecessary.

4. THEORETICAL BASIS OF THE THREE-COMPOSITE OPERATOR

We now examine the conditions for success and failure of the three-composite facial operator. In so doing no specific reference will be made to higher-order stages involved in facial processing; however, Bruce (1988) has noted the general utility of extracting three-dimensional shape primitives. On the basis of the experimental results reported in previous sections of this paper, it will be assumed that any early-vision operator which locates the rim of a three-dimensional solid (more strictly its projection on the image plane) provides powerful though not necessarily complete information about shape to the next-higher processing stage. The question is: how does it do this under varying conditions of illumination?

Consider the original postulate (figure 1) defining facial features. We note that *an image-space operator can detect a feature if and only if the luminance gradient along a line crossing the feature is non-zero.*

Imagine a vertical line drawn on the surface of a face as shown in figure 2*a*, starting at point A on the chin and ending at point C on the neck. The Gaussian map of this line is shown in figure 1. When the lower surface of the chin is reached at B, the surface normal will have turned through nearly 90° and the f-region entered (except for people with flat chins). The map of the line may subsequently disappear briefly onto the other side of the sphere if any part of the lower surface of the chin is occluded. It will reappear in reverse traverse back along the same path (figure 1) as the neck is reached. Thus there is doubling back or fold in the Gaussian map of the line.

If the facial illumination is from above, the underside of the chin and the upper part of the neck may be in shadow. In this case there is a gradual diminution in image luminance as the line is tracked downwards, followed by a step-function drop into shadow, as shown in graph (i), figure 3. The step-function drop could be picked up by an edge detector; equally, however, it could be located by a thresholding operation with an appropriate slicing level. With frontal illumination and no under-chin shadow the luminance may rise again on the neck, producing a valley at the f-region, as shown in graph (ii), figure 3. This will be located by a valley detector. Clearly therefore, the three-composite operator has a sufficient ensemble of detectors to pick out the chin line in a variety of illumination conditions. It is interesting that the artist who drew figure 2*a* has used the equivalent of a shading operator to outline the chin in one section (ABC) and a valley or edge operator in another (A′B′C′).

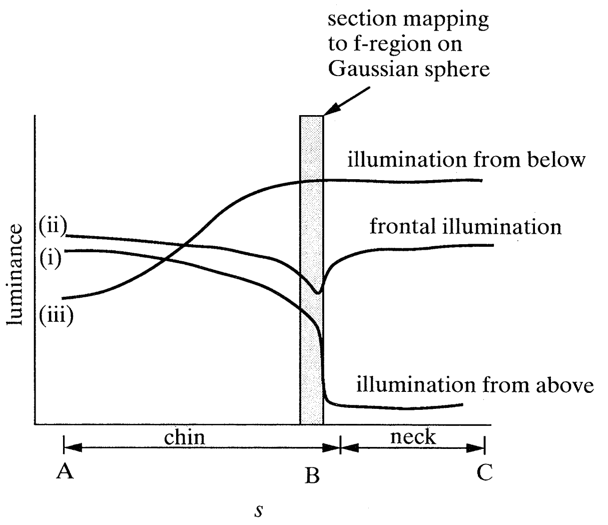

Figure 3. Image luminance as a function of distance *s* along the line ABC in figure 2, for various conditions of facial illumination.

He has also used shading with a sharp boundary to delineate the shape of the upper lip and the underside of the nose.

Note that a change of surface reflectance at the rim (as for example when the chin is seen in profile against a darker or lighter background) also produces a luminance step function (Pearson 1991). Gaussian-map representations are concerned with changes in surface normal, but it should not be forgotten that there are also changes in surface reflectance and texture.

Now consider a situation in which the illumination comes from below the face, so that the region under the chin is well illuminated. In this case it is possible that the luminance could gradually increase down the facial line, level off near the rim and not change much as the great circle is crossed, as in graph (iii), figure 3. Because there is no luminance gradient across the facial feature, neither an edge detector nor a valley detector nor a thresholding operation could locate it in image space.

5. WIDTH OF THE F-REGION AND PARABOLIC FACIAL CURVES

We now consider: (i) the occasional failures of the three-composite operator, as noted above; (ii) that some object boundaries, under certain conditions of illumination, produce image features which are not included in the ensemble of the three-composite operator, e.g. luminance ridges (Watt 1988); and (iii) that both human artist and three-composite operator draw lines to represent certain facial features (e.g. the sides of the nose, smile lines above the mouth, wrinkles) where there is not strictly a rim.

The importance of the parabolic curve or line has been emphasized by the work of Koenderink (1982, 1990). Parabolic curves divide synclastic surface regions, where the principal curvatures are of the same sign, from anticlastic regions, where they are of different signs; consequently at a parabolic curve one of the principal curvatures is zero.

The theory is advanced in this paper that only a subset of the parabolic curves on the face are of importance for any given illuminant and viewing angle. In particular, we refer to strong parabolic curves where the zero principal curvature increases rapidly with distance on at least one side of the curve, and to weak parabolic curves where this is not so. Thus strong parabolic curves divide strongly curved surfaces from neighbouring surfaces whose Gaussian curvature is of a different sign. Weak parabolic curves delimit weakly curved bumps, hollows or saddles. If a surface is strongly curved, then its luminance tends to change rapidly with distance along the surface in certain directions; with weakly curved surfaces the changes tend to be slow.

To understand the importance of parabolic facial curves we map surface luminance onto the Gaussian sphere, on the assumption that this luminance is determined primarily by the direction of the surface normal (this assumption obviously fails at times because of effects such as mutual illumination). The result is a reflectance map (Horn *et al.* 1989). We have made a few measurements of the reflectance maps of surfaces such as paper and (with difficulty) of human skin, in typical office illumination (Pearson 1990; Lee 1991). Figure 4 shows an amalgamation of two sets of measurements indicating very approximately the reflectance map for skin when the illumination is (*a*) from a point in front of the face and (*b*) from below the chin.

Superimposed on each reflectance map is a bold line which is the Gaussian map of the chin, as it might be represented in a line drawing. Coincidentally, the map has a chin-like shape! The head is assumed to be tilted back; this tilt causes all features on the face, including the chin line and the points A, B and C, to rotate upwards on the Gaussian sphere. The chin line is now a parabolic curve on the underside of the chin rather than a rim. It can be seen in figure 4*a* that the isoluminance curves run parallel to the chin line. Traversing the line ABC over the chin therefore produces both a fold in the map and a luminance minimum at B.

As the head is tilted further back, there comes a point when it is no longer appropriate to insert a chin line underneath the chin; rather the upper surface of the chin becomes more prominent and finally, as the face disappears from view, it becomes necessary to represent the rim on the top side. This can be handled in the theory by adjustment of the width of the f-region; however, no experimental evidence has as yet been gathered concerning its optimal width.

In figure 4*b* the illumination source is below the chin and the isoluminance contours cross the chin line nearly orthogonally, except at the point of the chin (B), where they are momentarily parallel. Traversing the line ABC now produces a luminance maximum at B. Such a situation requires the use of a ridge detector to extract the chin line.

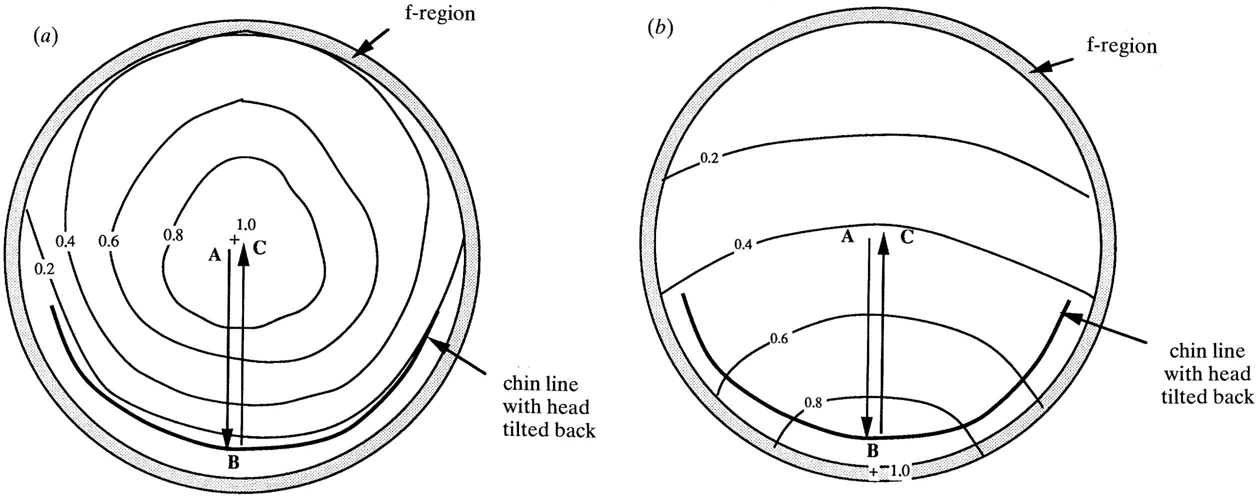
(a)
f-region
0.2
0.4
0.6
0.8
1.0
A
C
B
chin line with head tilted back
(b)
f-region
0.2
0.4
A
C
0.6
0.8
B
1.0
chin line with head tilted back

Figure 4. Approximate reflectance map for skin, showing isoluminance contours as a proportion of the peak luminance (1.0). Also indicated is the map of line ABC with the head tilted back. In (*a*) the illumination is directly in front of and in (*b*) from below the face.

6. DISCUSSION: ILLUMINANT DIRECTION AND CHOICE OF OPERATOR

It is apparent that under certain (relatively uncommon) conditions of illumination a ridge detector needs to be added to the armoury of a composite facial feature operator. This would make the existing three-composite detector into a four-composite detector, with an ability to detect valleys, ridges, edges and dark areas. In practice this might comprise, for the reason previously given, 'valledge', 'ridgedge' and thresholding operations.

It would be inadequate, however, merely to detect these features in parallel. A mixture of lines representing ridges and valleys on the face gives rise to a confusing picture (Haralick 1983). A four-composite detector would need to be linked to an illuminant-direction sensor and adaptively adjust its relative sensitivity to valleys and ridges. To appreciate this, consider figure 5, which portrays some hypothetical skin-surfaced feature bounded by parabolic curves, with the illumination coming from the side. In what follows we shall assume that the feature is oriented such that its centroid is near the centre of the diagram in figure 5. What this means is that if the feature is a nose, for example, then the nose points roughly towards the observation point.

We have noted previously that if a facial feature is strongly curved, then its Gaussian map extends over a large area. If the feature is a rather flat or gentle bump or hollow on the face, then its bounding parabolic curve is weak and the feature, if plotted, would map to a relatively small area near the centre of figure 5. Because the parabolic curves are weak, the luminance gradient at the feature boundary is small and difficult to detect. It is also difficult to see the feature in practice; for this reason it may not be important to represent it at all in a line drawing.

If on the other hand the feature is strongly curved, its Gaussian map will be large. In consequence, its bounding parabolic curves will map towards the extremities of figure 5, so that there will be an increased likelihood of their falling in the f-region. Because the principal curvature orthogonal to the feature boundary increases rapidly with distance from the boundary, there is the possibility that the luminance gradient across the boundary will also be large and therefore detectable. The feature actually represented in figure 5 is a strong feature i.e. it is bounded by strong parabolic curves. The f-region has been generously extended and the boundary of the feature (which does not have a rim) falls wholly within it.

In figure 5*a* the illumination highlight is at the extreme left of the diagram and outside the boundary. Consequently between points a and b the boundary generates a ridge in image space and between c and d it generates a valley. In between these points (from b to c from d to a) there is very little contrast across the boundary and it may be difficult in practice to detect it, even though it is bounded by a strong parabolic curve. In figure 5*b* the illumination source has moved so that the highlight now lies inside the map of the boundary of the object. Now both the ab and cd sections of the boundary are valleys. The sections bc and da are again of low contrast and difficult to detect.

These observations suggest the following rule. If there exists on a facial feature a surface normal whose direction corresponds to the highlight on its reflectance map, the f-region should be detected by a valley detector. If this is not the case, the f-region should be detected by the selective use of either a valley or a ridge detector; the ridge detector should be used on surfaces whose surface normal inclines towards the light source and the valley detector on surfaces which incline away from the light source. The boundary may be difficult to detect where isoluminance contours are nearly orthogonal to it.

If therefore a four-composite operator can be linked to a directional-illuminant sensor of sufficient precision to establish whether it falls inside or outside the Gaussian map of an object, then in principle the

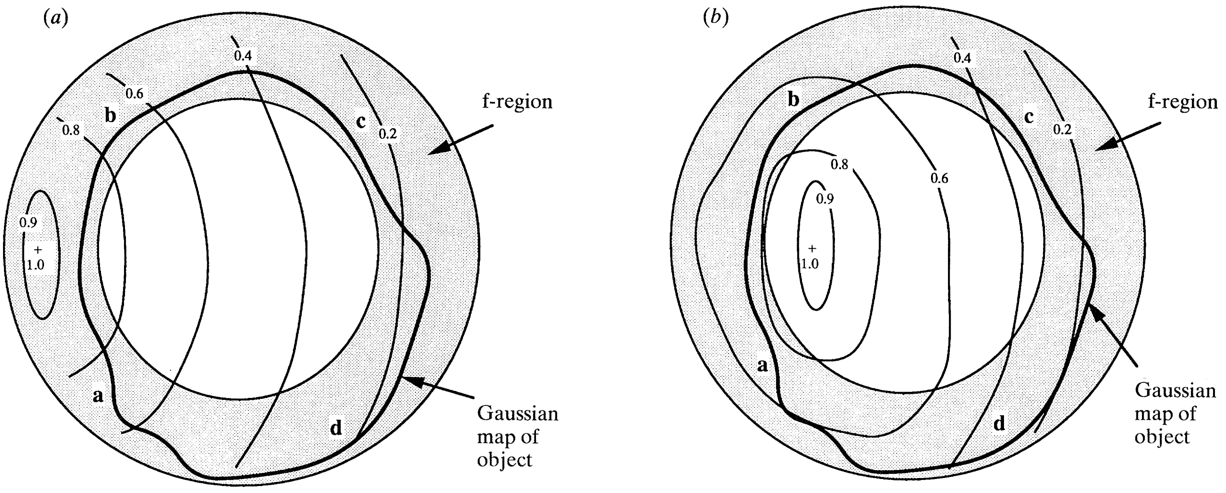

Figure 5. Isoluminance contours for hypothetical skin-covered feature bounded by parabolic curves (shown bold) with (*a*) illumination source well over to one side and (*b*) shifted to a point such that the maximum possible luminance (1.0) occurs on the object.

bounding surfaces of that object could be detected. Note, however, that this explanation confirms the observed result that the three-composite operator works well in a wide variety of frontal illuminants. Only when the illuminant is incident at a large angle to the viewing direction is there a need to use ridge detection. Note further that strong facial features tend to get detected by valley detectors because their Gaussian maps are large; with a large Gaussian map there is an increased chance, with a random direction of illumination, that a surface normal will exist which points towards this source. Weak parabolic curves, which tend to lie near the centre of the Gaussian map, are more likely to require ridge detection; however, the experimental evidence seems to be that both artists and machine tend to omit these weak curves in line drawings.

We are currently attempting to extend the rule to reflectance maps with more than one highlight.

7. CONCLUSIONS

A summary has been given of experimental work performed, over the last decade, into methods of extracting binary facial features for use in deaf communication systems. It has been found that economical and veridical line drawings of faces can be produced from facial images by a three-composite image operator having a sensitivity to valleys, edges and dark areas. This operator was used in the first-ever conversation by two deaf people over the public switched telephone network on 15 February 1989.

In the paper the theory that gave rise to the three-composite operator has been reviewed and extended. It has been shown that a promising basis for defining wholly skin-covered facial features is to map them onto a Gaussian sphere. The spatial extent of these features is represented on the face by either a rim or a strong parabolic curve; in practice these both tend to map into an annular region at the extremity of the Gaussian sphere, which has been called the feature region or f-region.

Conditions for being able to detect the bounding surfaces of a feature in image space have been put forward, with the area of the chin being taken as a particular example. It has been suggested that if there exists a highlight within the bounding curve of a feature, then a three-composite operator will suffice. However, when the illuminant direction forms an extreme angle with the viewing direction, a four-composite operator incorporating the adaptive use of a ridge detector may be necessary.

REFERENCES

Bergen, J.R. & Adelson, E.H. 1988 Early vision and texture perception. *Nature, Lond.* **333**, 363–364.

Bruce, V. 1988 *Recognising Faces*. London: Lawrence Erlbaum Associates.

Bruce, V., Hanna, E., Dench, N., Healy, P. & Burton, M. 1991 The importance of "mass" in line-drawings of faces. *Appl. Cogn. Psychol.* (In the press.)

Haralick, R.M. 1983 Ridges and valleys on digital images. *Comput. Vis. Graph. Image Process.* **22**, 28–38.

Horn, B.K.P. & Brooks, M.J. 1989 *Shape from Shading*. Cambridge, Massachusetts: MIT Press.

Koenderink, J.J. 1982 The shape of smooth objects and the way contours end. *Perception* **11**, 129–137.

Koenderink, J.J. 1990 *Solid Shape*. Cambridge, Massachusetts: MIT Press.

Lee, K.Y. 1991 Texture mapping in model-based image coding with luminance compensation. M.Sc. project report, University of Essex.

Letellier, P., Nadler, M. & Abramatic, J.-F. 1985 The Telsign project. *Proc. IEEE* **73**(4), 813–827.

Ono, H., Seki, H. & Deguchi, T. 1985 Animated TV telephone system for deaf people. *International Congress on the Education of the Deaf, University of Manchester.*

Pearson, D.E. & Sumner, J.P. 1976 An experimental visual telephone system for the deaf. *Jl R. telev. Soc.* **16**(2), 6–10.

Pearson, D.E. 1981 Visual communication systems for the deaf. *IEEE Trans. Comms.* **COM-29**(12), 1986–1992.

Pearson, D.E. & Six, H. 1982 Low data-rate moving-image transmission for deaf communication. *Proc. IEE Conference on Electronic Image Processing, York.* IEE Conference Publication no. 214, 204–208.

Pearson, D.E. 1983 Evaluation of feature-extracted images for deaf communication. *Electron. Lett.* **19**(16), 629–631.

Pearson, D.E. & Robinson, J.A. 1985 Visual communication at very low data rates. Proc. IEEE **73**(4), 795–811.

Pearson, D.E. 1986 Transmitting deaf sign language over the telecommunication network. *Br. J. Audiol.* **20**, 299–305.

Pearson, D.E. & Hanna, E. 1989 Operators for facial feature extraction. *Optical Society of America Topical Meeting on Applied Vision, San Francisco*. OSA Technical Digest Series **16**, 34–37.

Pearson, D., Hanna, E. & Martinez, K. 1990 Computer-generated cartoons. In *Images and understanding* (ed. H. Barlow, C. Blakemore & M. Weston-Smith) pp. 46–60. Cambridge University Press.

Pearson, D.E. 1990 Texture mapping in model-based image coding. *Image Commun.* **2**(4), 377–395.

Pearson, D.E. (ed.) 1991 *Image processing*. New York: McGraw-Hill.

Robinson, J.A. & Pearson, D.E. 1984 Visual teleconferencing at telephone data rates. *Proc. Int. Teleconference Symposium, London*. 359–364.

Robinson, J.A. 1985 *Low data-rate visual communication*. Ph.D. thesis, University of Essex.

Robinson, J.A. 1986 Low data-rate visual communication using cartoons: a comparison of data compression techniques. *Proc. IEE* part F, **133**(3), 236–256.

Sperling, G. 1981 Video transmission of American sign language and finger spelling: present and projected bandwidth requirements. *IEEE Trans. Comms.* **COM-29**(12), 1993–2002.

Sperling, G., Landy, M., Cohen, Y. & Pavel, M. 1985 Intelligible encoding of ASL image sequences at extremely low information rates. *Comput. Vis. Graph. Image Process.* **31**, 335–391.

Watt, R. 1988 *Visual processing: computational, psychophysical and cognitive research*, London: Lawrence Erlbaum Associates.

Welsh, W. 1991 Model-based image coding. Ph.D. thesis, University of Essex. (See also chapter in Pearson (1991).)

Whybray, M.W. & Hanna, E. 1989 A DSP based videophone for the hearing-impaired using valledge processed pictures. *Proc. IEEE International Conference of Acoustics, speech and Signal Processing (ICASSP.89), Glasgow.*

Whybray, M.W. 1991 Visual telecommunications for deaf people at 14.4 kbits/s on the public switched telephone network. *Cost 219 seminar on Videophony for the Handicapped, The Hague, Netherlands.*